FIRM
Formal Inference–based Recursive Modeling

Douglas M. Hawkins

Technical Report #546

April 1990

FIRM

Formal Inference-based Recursive Modeling

Douglas M Hawkins
Department of Applied Statistics
University of Minnesota
St Paul
MN 55108

Abstract

Recursive modeling is a largely model-assumption-free method of exploring the relationship between a dependent variable and a set of predictors. The data set is partitioned into two or more groups defined by ranges of values of one of the predictors. Each of the successor groups in turn is similarly partitioned into two or more groups defined by ranges of values of one of the predictors. The analysis continues until some termination rule indicates that none of the subgroups can be split further.

There have been a number of proposals for modeling based on recursive partitioning - notably Automatic Interaction Detection (AID), Classification and Regression Trees (CART) and Fast Algorithm for Classification Trees (FACT). The present code - FIRM differs from these in several respects - notably of varying the number of descendant nodes into which different nodes are split; and of using conservative formal (Neyman-Pearson) statistical inference for determining when to terminate analysis of each node.

Further differences include a facility for handing 'predictive missingness' - where the fact of a predictor's being missing conveys some predictive information about the dependent variable. The predictors are on either the nominal scale ('free' predictors) or the ordinal ('monotonic' and 'floating' predictors). The dependent variable may be on either categorical or on the interval scale of measurement - separate codes are provided for a categorical dependent variable (CATFIRM) and one on the interval scale (CONFIRM).

The codes run on personal computers. Executables are available for the IBM PC and the Apple Mac. Two versions of the IBM PC code are provided - a faster version which requires an 80286 board with math coprocessor, and a more general code which will run on a minimal PC. The Mac code has been tested on an Apple Mac Plus, but not on any smaller machine.

Introduction

The most common methods of investigating the relationship between a dependent variable Y and a set of predictors $X_1,,,X_p$ are the linear model for a continuous dependent variable (multiple regression, analysis of variance and analysis of covariance); and the log linear model for a categorical dependent. Both methods involve strong model assumptions. For example, the linear model assumes that the relationship between Y and the X_i is linear, and that there are no interaction terms other than those explicitly included in the model. The log linear model similarly assumes that interaction terms not explicitly included are zero, and while the LLM does provide (in the form of the omnibus test of fit) a check on whether all omitted interaction terms are zero, in the event that this test fails, it does not provide any direct indication of which interaction terms are needed.

A different approach is given by modeling based on recursive partitioning. In this approach, the calibration data set is successively split into ever smaller subsets, based on the values of the predictor variables. Each split is designed to separate the cases in the node being split into a set of successor groups which are in some sense maximally internally homogeneous.

An example of a data set in which FIRM is a potential method of analysis is the 'head injuries' data set of Titterington et al (1981). As we will be using this data set to illustrate the operation of the FIRM codes, and since the data set is included on the distribution diskette for testing purposes, it may be appropriate to say something about it. The data set was gathered in an attempt to predict from prognostic indicators the final outcome of patients who suffered head injuries. The outcome for each patient was that he or she was (i) dead or vegetative, (ii) had severe disabilities, or (iii) had a good or moderate recovery. This outcome is to be predicted on the basis of 6 available predictors:-

1 Age. The age of the patient. This was grouped into decades in the original data, and is grouped the same way here. It has 8 classes.

2 EMV. This is a composite score of three measures - of eye opening in response to stimulation; motor response of best limb; and verbal response. This has 7 classes, but is not measured in all cases, so that there are 8 possible codes for this score - the 7 measurements and an eighth 'missing' category.

3 MRP. This is a composite score of motor responses in all four limbs. This also has 7 measurement classes with an eighth class for missing information.

4 Change. The change in neurological function over the first 24 hours. This was graded 1, 2 or 3, with a fourth class for missing information.

5 Eye indicator. A summary of diagnostics on the eyes. This too had three measurement classes, with a fourth for missing information.

6 Pupils. Pupil reaction to light - present, absent, or missing.

Figure 1 is a dendrogram showing the end result of analyzing the data set using the CATFIRM code, which is appropriate for a categorical dependent variable. nd Figure 2 (for a continuous dependent variable). In Figure 1, we see that the most significant separation was obtained by splitting the full sample ('node number 1') into two on the basis of the predictor 'Pupils'. Cases for which Pupils had the value 2 or the value ? (ie missing) constitute one of the successor groups ('node number 2'), while those for which Pupils had the value 1 constitute the other ('node number 3'). Then each of these nodes in turn is subjected to the same analysis. No way can be found to split the cases in node number 3 any further, so this node is 'terminal'. The cases in node number 2 however can be split into more homogeneous subgroups. The most significant such split is obtained by separating the cases into four groups on the basis of the predictor 'age'. These are patients under 20 years old, (node 4), patients 20 to 40 years old (node 5), those 40 to 60 years old (node 6) and those over 60 (node 7).

These groups, and their descendants, are analyzed in turn in the same way. Ultimately no further splits can be made. Altogether 17 nodes are formed, of which 11 are terminal.

The dendrogram and the analysis giving rise to it may be used for predictive purposes, or for further understanding of the importance of and interrelationships between the different predictors. Taking the prediction use first, the dendrogram provides a quick and convenient way of predicting the outcome for a patient - finding out into which terminal node a patient falls yields 11 typical patient profiles ranging from 90% dead/vegetative to 86% with moderate to good recoveries. The fuller analysis shows that all 6 predictors are very discriminating in the full data set, but are much less so as soon as the initial split has ben performed, diagnosing a high degree of commonality in their predictive information. Another feature often seen (though not very strongly in this data set) is an interaction in

which one predictor is predictive in one node, and a different one is predictive in another. Examples of this phenomenon are given in Hawkins and Kass (1982).

The other analysis covered by the FIRM package is for a dependent variable on the interval scale of measurement - this is the CONFIRM code. Figure 2 shows an analysis of the same data using CONFIRM. The dependent variable was on a three-point ordinal scale, and for the CONFIRM analysis this was regarded as being on the interval scale of measurement - an assumption that is not necessarily tenable, but which we make for the convenience of illustrating both codes with a single data set. Here, the first split is made on the basis of the predictor MRP. Cases for which MRP is 1 or 2 constitute the first descendant group ('node number 2'), those for which it is 3, 4 or 5 give 'node number 3', while those for which it is ? (ie missing) 6 or 7 give 'node number 4'. The cases in node number 2 are then found to be capable of being split again. The predictor Pupils gives the most significant split, defining 'node number 5', cases for which Pupils is either ? or 1, and 'node number 6', - the cases for which Pupils is 2. Pupils is also used to split node 3, but a different predictor - Age is used for node 4.

Continuing down, node 7 shows a feature sometimes found - an outlier. Among the 54 cases, nearly all of whom ended up dead or vegetative, was a single case with EMV=6 who made a good recovery. FIRM handles outliers by isolation, stripping out small groupings such as that seen here.

In this analysis, a total of 17 nodes are formed, of which 10 are terminal.

There are three elements to the splitting by FIRM (or any other recursive partitioning algorithm) - (i) deciding which predictor to use to define the split; (ii) deciding which categories of the predictor should be grouped together so that the data set is not split more ways than are really necessary (ie is a binary, threeway, fourwayy ... split on this variable required); and (iii) deciding when to stop growing the tree. Different implementations of recursive partitioning handle these questions in different ways..

The dendrograms of Figures 1 and 2 were produced manually from the detailed output, a sample of which is given in appendices. This output gives much useful information:-

1 For each split, which predictor produces the split, which categories of the predictor are grouped together, and what the conservative statistical significance of the split is;

2 The number of cases flowing into each of the descendant groups;

3 Summary statistics of the cases in the descendant nodes. In the case of CATFIRM, the summary statistics are a percentage frequency breakdown of the cases between the different classes of the dependent variable. With CONFIRM, the summary statistics given are the arithmetic mean and standard deviation of the cases in the node.

The theoretical basis for the procedures implemented in FIRM and used to produce these dendrograms is set out in detail in Hawkins and Kass (1982) and Kass (1980). Antecedents of the CATFIRM and CONFIRM codes are discussed in Kass (1975) and Heymann (1981) respectively. The interested reader should refer to this exposition and to the papers referenced there for detail of the methodology used.

Terminology

FIRM comprises two codes, one for a categorical dependent variable and the other for an interval dependent variable. These are called the CATFIRM and CONFIRM codes respectively. Both require categorical predictors. The maximum number of categories that can be accommodated is set at compilation time, but is in the region of 15 to 20.

If a predictor is on the interval scale, it must first be grouped into distinct classes before it can be analyzed by FIRM. On the face of it, this is a serious limitation of the procedure, but on closer inspection this is seen not to be the case. The end result of the FIRM analysis is itself a grouping, and all that the preliminary need to group the predictor does is to place some restrictions on the places that a split can occur. For example, consider a continuous predictor whose values range from 0 to 100. For FIRM analysis, this might be grouped into classes - say 0-10, 10-20, 10-30, ... 90-100. Then in the FIRM analysis, only multiples of 10 would be eligible split points whereas an analysis which did not have this limitation would be able to use any value as a split point. As it is seldom the case that a split at say 55 would be a great deal better than splits at either 50 or 60, the limitation of the split points is seldom a source of much loss in performance.

This grouping is exemplified in the head injuries data. Age is a continuous variable. However both in the original paper and our modeling, age has been grouped into decades. This means that only ages that are multiples of 10 years may be used as split points in the analysis, leaving the possibility that, when age was used to split node 3 in CATFIRM with cut points at ages 20, 40 and 60, better fits might have been obtained all ages inbetween the decade anniversaries been allowed as possible cu points. In a data set such as this however, it seems unlikely that there would be any substantial extra benefit of being able to split at these intermediate ages.

While all categorical, there are three subtypes of predictor:-

(i) FREE predictors are on the nominal scale. When grouping categories together, any classes of the predictor may be grouped.

(ii) MONOTONIC predictors are on the ordinal scale. When the classes of a predictor are considered for grouping, only groupings of contiguous classes are allowed. For example, in an 8 class monotonic predictor, it would be permissible to pool into the groupings {1,2}, {3}, {4,5,6,7,8}, but not into (1,2),{5,6),{3,4,7,8}. The term 'monotonic' refers to the expectation that the response would be monotonic in the predictor, and that accordingly the FIRM step function would be a suitable (approximate) model. In some circumstances, a predictor may be on the ordinal scale, but with no expectation that the response would be even smoothly dependent on it. In such cases, it would be sensible to regard the predictor as free and not monotonic.

(iii) A variant of the monotonic predictor is the FLOATING predictor. A floating predictor is one whose scale is monotonic except for a single 'floating' class whose position in the monotonic scale is unknown. (In the present implementations of FIRM, this class is required to be the first class). The rules of grouping are that the monotonic portion of the scale can be grouped only monotonically, but that the floating class may be grouped with any other class or classes on the scale.

As the 'floating' predictor type is not supported by most other recursive modeling procedures, some comments on its motivation and potential may be in order. The floating predictor type is particularly useful for handling missing information in an otherwise ordinal predictor.

Most models for missing information have some sort of underlying 'missing at random' assumption, but this is not the case with FIRM's floating category. Here, the fact that a predictor is missing may be informative about the dependent variable, and if this is the case, it will be diagnosed by the grouping that emerges - either the floating class will be isolated from all other classes, or it will be merged with classes toward one end of the scale. If the predictor really is missing at random on the other hand, then the floating category will be found grouped with a class in the center of the predictor's scale.

There are many examples of potentially informative missingness; we mention two. In the head injuries data, observations on the patient's eye function could not be made if the eye had been destroyed in the head injury, which will be more common with severe injuries than with mild. Thus it is not reasonable to assume that the eye measurements are in any sense missing at random; rather it is a distinct possibility that missingness could predict a poor outcome. Inspection of the groupings produced by FIRM does indeed suggest informative missingness on several predictors; EMV and MRP, with their 7-class monotonic plus floating class display this quite clearly when the missing class is grouped with classes 6 and 7 on the scale.

Another example we have seen is in educational data, where we attempted to predict college statistics grades on the basis of a long list of predictors, among which were high school math scores. In the student pool under investigation, high school math was not a prerequisite for college entry, and students who had not studied math in high school therefore had missing values for their high school math grade. However since it was often the academically weaker students who elected not to study math in high school, students missing this grade had below average low success rates in college. FIRM diagnosed and treated this automatically by grouping students with missing high schol math scores with students having the lowest passing grades in their high school math.

No special predictor type is needed for missing information on a free predictor; all that is necessary is to have a specific class for the missing values. For example, if in a survey of adolescents one were measuring family type in four classes:- 1: with both original parents; 2: with single parent; 3: with blended family; 4: with adoptive family; then respondents for whom the family type was not observed would define a fifth class, and the five classes would define a nominal scale.

In the head injuries data set, age was always observed, and is clearly on the MONOTONIC scale. EMV, MRP, Change and Eye indicator all have a monotonic scale but with missing information. Therefore they are used as FLOATING predictors. With Pupils, we have a choice. The outcome was binary with missing information making a third category. This situation can be handled equivalently as a floating or as a free predictor. Since there is no free predictor in the problem we will treat it as FREE and use it to illustrate the processing of free predictors.

With floating and the free predictors having a 'missing' category, it is an interesting piece of follow-up analysis to see with which (if any) of the base groups the 'missing' category is grouped by FIRM. Apart from giving structural information about the relationship between the missingness and the dependent variable, this may provide a useful missing information predictor. For example, if the FIRM analysis groups the missing information category with say categories 4, 5 and 6 in a floating scale, then where it is desirable to fill in some sensible value for the missing information, the middle class of the grouping (ie class 5) might be used.

Operation

Both CONFIRM and CATFIRM follow the same overall approach. At each node, the cases are analyzed using each predictor in turn. If the predictor has c classes, the cases are first split into c separate groups corresponding to these classes. Then tests are carried out to see whether these classes can be reduced to fewer classes by pooling classes pairwise. This is done by finding two-sample test statistics (Student's t for CONFIRM, chi-squared for CATFIRM) between each legally poolable pair of classes. If the most similar pair fail to be significantly different at the user-selected significance level, then the two classes are merged into one composite class. The pairwise tests are then repeated for the reduced set of c-1 classes. This process continues until no legally poolable pair of simple or composite classes is separated by a non-significant test statistic, ending the 'merge' phase of the analysis. Next, to protect against occasional bad groupings formed by this incremental approach, FIRM tests each composite classes to see whether it can be resplit into two that are significantly different at the 'split' significance level set for the run. If this occurs, then the composite group is split and FIRM repeats the merging tests for the new set of classes.

The end result of this repeated 'Fisher-LSD' type of testing is a grouping of cases by the predictor. All classes may be pooled into a single one, indicating that the predictor has no descriptive power in that node. If this does not occur then the analysis ends with from 2 to c composite groups of classes, with no further merging or splitting possible without violating the significance levels set.

FIRM's method of reducing the classes of the predictors differs from approaches in which the number of descendant nodes is fixed. Most other recursive partitioning procedures fix the number of ways a node splits - commonly considering only binary splits.

The final part of the FIRM analysis of a particular predictor is to associate with it a formal significance level. In doing this, it is essential to use significance levels that reflect the grouping of categories that has occurred between the original c-way split and the final say k-way split. Hawkins and Kass (1982) mention two ways of measuring the overall significance of a predictor conservatively - the 'Bonferroni' approach, and the 'Multiple comparison' approach. Both compute an overall test statistic of the k-way classification: for CATFIRM a Pearson chi-squared statistic, and for CONFIRM a one-way analysis of variance. The Bonferroni approach takes the P-value of the resulting test statistic and multiplies it by the number of implicit tests in the grouping from c categories to k. The multiple comparison approach computes the P value of the final grouping as if it had been based on a c-way classification of the cases. Since both approaches yield a conservative bound on the significance, the smaller of the two values is taken to be the overall significance of the predictor.

The final stage is the decision of whether to split the node further, and if so, using which predictor. This is done by finding which predictor is most significant on the conservative test, and making the split if its conservative significance level meets the user-selected cutoff.

The FIRM analysis stops when none of the nodes has a significant split.

Some users in some circumstances prefer, regardless of statistical significance, not to further split any nodes that are very small, or whose members are very homogeneous. Others wish to limit the analysis to a set maximum number of nodes. The codes contain options for these

preferences, allowing the user to specify threshold sizes (both codes) and a threshold total sum of squares (CONFIRM) that a node must have to be considered for further splitting, and to set a maximum number of nodes to be analyzed.

Comparison with other codes

Before discussing the details of using the codes, we mention some other related approaches. Notable early work in the area was that of Morgan and Sonquist (1963), who defined the 'Automatic Interaction Detector' (AID). This covered monotonic and free predictors, and made only binary splits. At each node, the split was made whose explained sum of squares was greatest, and the analysis terminated, not on the basis of any formal procedure, but when all remaining nodes had total sums of squares below some threshold.

Users found that the lack of a formal basis for stopping made the procedure very prone to overfitting. The use of explained sum of squares without any modification for the number of starting classes or the freedom to pool them also made AID tend to prefer predictors with many categories to those with fewer, and to prefer free predictors to monotonic.

Breiman et al (1984) defined the Classification and Regression Trees (CART) approach. This has two classes of predictors - categorical (corresponding to our 'free' predictors) and continuous. Continuous predictors are like our 'monotonic' predictors, but without our requirement that continuous predictors first be reduced to a smaller number of classes. CART does not have an equivalent of our 'floating' category. Instead missing values of predictors are handled by 'surrogate splits', by which when a predictor is missing on some case, other predictors are used in its stead. This approach does not seem to be as effective as our explicit provision for predictive missingness as it depends on the predictive power of the missingness being captured in other non-missing predictors.

CART uses only binary splits. Where FIRM chooses between different predictors on the basis of a formal test statistic for identity, CART uses a measure of 'node purity', with the predictor giving the purest descendant nodes being the one considered for use in splitting. Like AID, the node purity measure tends to prefer categorical predictors to continuous, and to prefer predictors with many distinct values to predictors with few.

With a continuous dependent variable, CART's measure of node purity comes down to essentially the same measure as a two-sample t test (like FIRM's), but the purity measure for a categorical dependent is substantially different. Here CART will attempt to find a two-column cross classification in which different categories of the dependent variable are modal in the different columns. FIRM has no such objective, but simply tries to find groupings that give large chi squared values. This has implications for the use of the two procedures in minimal-model classification problems. A FIRM analysis may produce very structured, highly significant splits, but end with a set of terminal nodes all of which are modal for the same class of the dependent variable. Using this dendrogram for minimal-model discriminant analysis will, unless some steps are taken to prevent it (like altering prior probabilities of the classes) lead to all unknowns being classified to the same class. This tendency is less strong with CART.

Two major differences between CART and FIRM relates to the rules that are used to decide on the final size of tree. FIRM creates its trees by 'forward selection'. This means that as soon as the algorithm is unable to find a node which can be split in a statistically significant way, the analysis stops. Since it is possible for a non-explanatory split to be needed before a lower-level explanatory one can be found, this means that it is possible for the FIRM analysis to stop too soon and fail to find all the explanatory power in the predictors. CART uses the opposite strategy of 'backward elimination'. First, a deliberately oversized tree is created by continuing to split nodes whether the splits produce some immediate benefit or not. Then the oversized tree is pruned from the bottom, undoing splits that appear not to correspond to genuinely distinct patterns in the data.

The second difference between CART and FIRM is more profound - it is in the procedure used to decide whether a particular split is real, or simply a result of random sampling fluctuations giving the appearance of an explanatory split. FIRM addresses this question using the Neyman Pearson approach to hypothesis testing. At each node, there is a null hypothesis that the node is indivisible; and this hypothesis is tested using a controlled conservative significance level. This emphasis on Type I errors should lead to many Type II errors, and a further tendency for FIRM to produce trees smaller than they should be. CART decides on the value of a split using cross validation. A record is kept of the actual trees formed with a total of say M nodes. Then a repeated cross validation is made. In this, the sample is randomly split into a larger 'calibration' portion and a smaller 'test' portion. The

'calibration' portion is used to generate a splitting with M nodes, and the hold-back 'test' cases are run through the tree and the node purity evaluated for the full tree and all subtrees. The sample is then randomly resplit into a different 'calibration' and a 'test' portion, and the calibration portion used to create yet another M node tree, whose subtrees are again evaluated by classifying the 'test' data and evaluating node purity. Many (typically 10) such trees are constructed. They need not resemble each other at all, though one would commonly hope that they did, at least for the higher levels. In this way, arguing by the analogy of different trees of the same size, a picture is built up of the node purity of generic trees of 1, 2, 3, ..., M nodes. The number of nodes which appears to give the greates node purity is selected as the correct tree size. The original tree of the full data set is then pruned to that number of nodes. Breiman et al point out the attraction of this measure of node purity - that it is 'honest', unlike that obtained by resubstituting the same data used to calibrate a tree.

Typically a CART analysis will produce 11 M-node trees - one of the full data set and another 10 from a ten-fold repeat of the cross validation by a calibration and test subsamples. This, together with the much greater initial size of tree, accounts for CART having much longer execution times than FIRM.

In other types of analysis (for example multiple linear regression) it is commonly observed that cross validation provides results quite like those using formal significance tests, but at a high significance level (for example 15% to 25%). This experience gives yet another reason to expect that FIRM with its testing at much smaller significance levels would give smaller final trees than does CART. This however appears not to be the case; in many analyses of different data sets, the FIRM trees have been found to be generally bigger and more detailed than the CART trees. This difference seems to be biggest where the association between the predictors and the dependent variable is relatively weak, though highly significant, and smaller in data sets (like the head injury data) where the association is strong. The CART tree for the head injury data, for example, has 15 nodes of which 8 are terminal - 2 fewer than FIRM.

We have carried out several trials using large data sets in which the data set is split in half; with one half analyzed by FIRM to calibrate a model which is then verified using the other half. As one would expect of a method based on controlled small significance level, these

trials have shown that the detail obtained using FIRM is real.

It thus appears that while the FIRM trees are themselves almost certainly smaller than they should be for optimal explanatory power, those of CART are smaller yet.

Another recursive partitioning method is the Fast Algorithm for Classification Trees (FACT), developed by Loh and Vanichsetakul (1988). This method was designed for a categorical dependent variable, and treats the splitting by using a discriminant analysis paradigm. The sample is split, not by predictor variables, but by the dependent variable. By an inverse regression analogy, the variable (or linear combination of variables) that is best separated by the dependent variable classes is taken to be the variable that best predicts the dependent variable. Continuous dependent variables are handled by discretization into a number of classes.

Running FIRM

The FIRM distribution package contains three FORTRAN executable files. Two of these, CONFIRM and CATFIRM carry out the recursive analysis for a continuous and a categorical dependent variable respectively. Both require details of the dependent and predictor variables as well as run options, and these details are specified in a control deck. The third program, DECK, is an interactive interface to create these control decks, and is the most convenient way for most users to do the setup necessary to run FIRM analyses. Regular users who prefer to change the options in the deck with a text editor should have no trouble interpreting the structure of the deck.

DECK first asks for the name of the file into which it must write the control deck created.

Next it asks whether you wish to modify an existing deck or create a new one. If you have already made a FIRM run with a particular data set and want to change some details of the run, it will generally be more convenient to use this option to modify the old deck than to create a new one.

If you do select the deck modification option, DECK asks for the name of the file that contains the existing deck. You may give the same file name for both the old and the new

deck; DECK will then write the new deck in the place of the old deck. With the deck modification option, DECK will prompt you at three stages of the run asking whether you want (i) to make any changes to the details given for the dependent variable; (ii) to make any changes to information about the predictors; and (iii) to change the run options. If you select any of these options, you will get the same series of prompts for that section as you do when creating a new deck; if you do not, the information on that section is copied from the old deck over to the new.

Dependent variable section The first set of substantive questions relates to the dependent variable. The first is where the dependent variable is in the data. For each case, a number of data values are to be read. Any of these may be the dependent variable, and this first question ascertains which of them it is. For example if the data file contains for each case three predictors, followed by one variable you do not want to use in the FIRM analysis, followed by the dependent variable, and then another four predictors, the answer to this first question is 5 (ie the dependent variable is the fifth variable read for each case).

Next DECK asks for the name of the dependent variable. Enter a name of up to 50 characters.

Then DECK asks how many categories the dependent variable has. If the dependent variable is categorical (so that you will be analyzing it using CATFIRM), then type the number of different categories it has. If the dependent variable is continuous (so that you will be analyzing it using CONFIRM), then type zero to this question.

Finally, if you said that the dependent variable had 1 or more categories, DECK will ask you for names for each category. Give the categories names of no more than 20 characters.

This ends the dependent variable section.

Predictor section The next section of DECK asks for details about the predictor variables. If you are modifying an existing deck, then DECK will ask for the numbers of the predictors you want to change, and will accept new details for those predictors only. Otherwise, you will be prompted for all predictors. The first question is how many predictors there are. Enter the number of predictors. Then DECK loops through for each predictor, asking for the

following pieces of information:-

1 The position of the predictor in the data. Returning to our earlier example in which the data file contained for each case three predictors, one variable you do not want to use, the dependent variable, and then another four predictors, there are a total of 7 predictors. Their positions in the data are 1, 2, 3, 6, 7, 8 and 9.

2 The smallest value of the predictor. FIRM requires that predictors take on consecutive integer values; this question asks what the smallest of the values is. Commonly the values will run from 1 up, or from 0 up, but any other smallest value can be handled.

3 The type of the predictor. Enter m if the predictor is monotonic; f if it is free, and 1 (ie the digit 'one') if it is a floating predictor, the smallest value of whose scale is used to indicate that the predictor is missing on that case.

4 Information on whether the variable may be used for splitting or not. Commonly, all predictors are usable - ie if any predictor gives the most explanatory split, then the data will be split on that predictor. On occasion though there will be potential predictors which you do not want to use for splitting. Such predictors may either be omitted from the list of predictors (like the fourth variable in the example data file mentioned above), or included, but flagged unusable. FIRM will carry out the same analysis for unusable as for usable predictors, showing how the categories of the predictor are grouped and providing the significance of the split on the unusable predictor, but that predictor will not be used for splitting. This option is useful in a number of circumstances - for example an otherwise good predictor might be difficult to measure, so one might be interested to know whether it retained significance when other predictors were used in its place. This can be done by carrying the predictor in the analysis but making it unusable, and checking the details of its analysis and significance levels in the successive splits.

4 The number of categories. A predictor taking values 0, 1, 2, 3 for example has 4 categories.

5 Category symbol list. This option calls for a one-character label to be associated with each category of a predictor. For example if a predictor has categories corresponded to missing, poor, fair, good and excellent, then it is mnemonic to label these categories ?, p, f, g, and e respectively. In responding to this query, list the symbols with no intervening spaces - ie type in ?pfge

6 Split/merge significance levels. FIRM operates by carrying out sequences of two-sample tests to group the categories of each predictor. This query asks for the significance level to be used for the test to split a composite category; and to merge a pair of (simple or composite) categories respectively. If the split significance level is set higher than the merge significance level, then no splitting will be carried out; only merging. This can lead to some saving in execution time in CATFIRM for free predictors with many categories, but is not generally advisable. The significance levels are given in percent. For example, if you wish to test at the 0.9% significance level for splitting and the 1% for merging, then enter the values 0.9 1. Common choices for these significances are values close to 5%. If the significance levels are set to smaller values, then there will tend to be fewer final categories on any predictor since a more stringent test will be applied for keeping categories separate.

Run options The final set of queries from DECK sets run options. All of these have defaults, which (for numeric queries) are obtained by giving the value 0 in response to the query. The 'dependent variable' and the 'predictor variables' section of DECK are the same for CATFIRM and CONFIRM, but the options are different, so it is necessary to discuss these separately.

CATFIRM run options

1 The first option relates to the printing of the contingency tables generated in the solution. You may print these in 4 possible formats - (i) as percentages of the column totals, (ii) as overall percentages of the grand total, (iii) as percentages of the row totals, and (iv) as raw counts. CATFIRM gives contingency tables with the dependent variable forming the rows, and the predictor the columns. Thus the first and generally most useful option (to get which enter 0) gives the table as a percentage frequency breakdown of the dependent variable for all cases in the node, and for the cases broken down by the different categories of the predictor. The last line of each contingency table is the total number of cases at that value of the predictor. The second option (enter 1) gives the individual cell frequencies, and the row and column totals, as percentages of the grand total frequency in that table. The third (enter 2) produces a percentage frequency breakdown of the different

categories of the predictor for each level of the dependent variable, while the fourth (enter 3) gives the actual frequencies in each cell, and the marginal totals, of the contingency table.

2 'Do you want details of splits not used'. Both FIRM programs will, on request, produce details of the analysis of each predictor in each node, showing the successive steps in the splitting and merging that produces the final grouping of each predictor, and giving the test statistics at each stage of the grouping. This output can be very useful, but is voluminous, particularly where many nodes are created, where there are many predictors, and/or the predictors have many levels. This option allows you to request, or to suppress the gory detail. If the detail is requested, it will be put on a separate file - the 'detailed split' file. Whether you request this detailed output or not, FIRM will give an overall summary of each predictor in each node, showing how its categories group and what the final significance of the split is.

3 CATFIRM will optionally produce a separate file of the contingency tables associated with each predictor in each node. Like the detailed split output, this file while often useful can be very big, and so CATFIRM provides the option of limiting or suppressing it. Two questions are asked about cross tabulations - (i) whether you want to see the cross tabs before grouping, and (ii) whether you want to see them after grouping. If both are selected, then the 'Tables' file will contain for each predictor in each node, the full cross tabulation before the grouping of the predictor's categories takes place, and again for the finally grouped categories.

4 The next query is of the minimum size a node must have to be considered for further splitting. There are two reasons for using this option. One is that the chi-squared approximation to Pearson's X^2 statistic deteriorates when frequencies in the contingency table get small, and the other is that in practical terms one might not be interested in finding out about the splits possible in nodes containing only few cases. If no value is set for this option, the default value of 50 is used.

5 Two questions are asked - the minimum raw significance, and the minimum conservative significance level that a split must attain to be used. When the analysis of a predictor is complete, the result is an RxC contingency table, the C representing the number of composite categories after grouping. Associated with this table is a Pearson X^2 value. The raw significance level is obtained by entering X^2 in a χ^2 table with (R-1)(C-1) degrees of freedom. Setting the first of these significance

levels to say α means that the raw significance level must attain at least α for the split to be made. The conservative significance level is a more realistic value that takes into account the effect on X^2 of the grouping that occurs in analyzing the predictor. Setting the second significance level to say α means that the conservative significance must attain at least α% for the split to be made. Generally, the testing will be driven by the second, and not the first of these significance levels; since the conservative P value is always at least as large as the raw, setting the two values equal ensures this. The default for both significance levels is 5%. Smaller values than this should be used when there are many predictors. A conservative choice would be 5/M, if M is the number of predictors in the study.

6 Next you are asked the maximum number of groups to analyze. Analysis will terminate when this many nodes have been investigated for splitting. This limit applies to the number of nodes analyzed' the number formed may exceed this, the excess nodes being on the unresolved list when FIRM terminates.-

7 Entering a nonzero constant A for the next option causes CATFIRM to compute, instead of X^2, a statistic whose summand is (observed - expected)^2 ÷ (expected + A). Entering the usual choice of zero gives the standard Pearson X^2.

8 Finally, DECK asks whether the data are in free or fixed format. Free format is more convenient and is possible when (i) the data file contains only numeric values, and (ii) all values are separated by at least once blank. Fixed format is required if either of these conditions fails. If the data have to be read in fixed format, DECK next asks for a FORTRAN format specification in which to read the data. CATFIRM's data have to be read with INTEGER specifications. If it is necessary to use fixed format and the data file contains variables that are neither the dependent variable nor predictors to be analyzed, then it will be most efficient to give a format specification that skips over these values. When this is done, the 'position in the data' needed for both dependent and predictor variables refers to the position amongst those variables actually read.

This completes the list of queries for a CATFIRM run. After DECK has executed, the control deck will be ready for feeding to CATFIRM.

CONFIRM run options

CONFIRM uses many of the same options as CATFIRM, but not all are implemented quite the same, or the same order. The list of the queries from DECK is

1 Whether you want the detailed split file. Like CATFIRM, CONFIRM is able to produce the details of the analysis of each predictor at each node, giving the values of the two-sample t statistics used to decide whether to merge (simple or composite) groups. While not as large as the corresponding CATFIRM output, this file can still be large, and frequently users will suppress it.

2 CONFIRM allows the user to set thresholds on both the size and variability a node must have to be considered for splitting. Using these thresholds prevents study of very small or very homogeneous nodes. The next two questions from DECK are for the minimum size a node must have, and for the minimum proportion of the initial sum of squared deviations from the mean it must have, to be considered for analysis. The first of these options is clear, to illustrate the second, suppose one set a value of 0.001. Then any node whose sum of squared deviations from the mean fell below 0.001 times that of the original full sample would not be considered for splitting.

3 Next DECK calls for the raw significance level, and conservative significance levels that a predictor must attain for the split to be made. Default values for both are 5%.

4 DECK then calls for the maximum number of nodes to be analyzed. When this number of nodes have been investigated for splitting, the analysis terminates. The number formed may exceed this maximum, the excess number being nodes that have been formed, but not yet analyzed.

5 Next, DECK asks whether the data are in free format or fixed format. As with CATFIRM, provided the data file (i) contains only numeric data, and (ii) has at least one space between each pair of data values, free format will generally be the most convenient way of reading the data. If either of these conditions fails however, it will be necessary to read the data in fixed format. If this is the case, DECK asks for the FORTRAN format specification for reading the data.

There is an important incompatibility between CONFIRM and CATFIRM when fixed format is used - while CATFIRM requires INTEGER specifications for reading the data, CONFIRM (one of whose variables is continuous) requires REAL specifications for all variables - dependent and predictor.

6 The last option requested by DECK relates to the denominator variance used for the t tests. The first stage of the analysis of a predictor is to form the count, mean values, and sum of squared deviations from the mean of the cases with each value of the predictor variable. These numbers can be used to form a one-way analysis of variance table. When subsequently testing pairs of categories for compatibility, there are two natural candidates for the variance term in the denominator of the t statistic - the pooled variance of just those two groups being tested, and the pooled error variance of the one-way analysis of variance. You are asked to select one of these. Neither is uniformly superior to the other - the pooled variance brings more information to bear on the test (which is good), but if the data contain outliers or heteroscedasticity, then pooling may contaminate the good information for a particular pair of categories with bad information from other categories (which is bad).

Once these options have been supplied, DECK terminates with a message that you are ready to run CONFIRM.

Where the dependent variable is ordinal, and there is reason to suppose that is may be close to interval, you may want to analyze the same data set with both CATFIRM and CONFIRM. Sometimes the opposite holds - you may have an interval dependent variable but be interested in establishing whether its distribution (and not just its mean) depends on the predictors. This can be done by running with CONFIRM, then discretizing and running with CATFIRM. It is permissible to use DECK to modify a CONFIRM deck to control a CATFIRM run and vice versa. When doing this, you MUST respecify the dependent variable section (because the two modes of analysis differ in the number of categories specified for the dependent variable); you will generally leave the predictor section unchanged; and (unless using default values throughout both runs), will generally want to respecify the run options.

Once the control deck has been created, you are ready to run CATFIRM or CONFIRM. This is done by starting CATFIRM or CONFIRM in the usual way (type CATFIRM or CONFIRM on the IBM; double click CATFIRM or CONFIRM on the Mac). When this is done, you are prompted to enter certain file names from the keyboard, and when these are supplied the FIRM analysis proceeds.

CATFIRM calls for the names of 6 files. These are:-

1 'Control deck'. This is the file containing the control deck created using DECK.

2 'Data'. This is the file containing the data for the analysis. These first two are input files.

3 'Summary'. This is the primary output file created by CATFIRM. It contains a section for each node created in the analysis showing (i) a list of all predictors and the grouping made of the categories of each; (ii) the raw, multiple comparison, and Bonferroni significance of a split on that predictor; (iii) which predictor (if any) is used for the split, and (iv) the (grouped) contingency table for of the dependent variable against the predictor used for the split.

4 'Split'. This output file from CATFIRM contains the optional full details of the test statistics used in the reduction of each predictor at each node to its final grouping. It is necessary to give a file name for the detailed splits even if the control deck does not call for the detailed split output.

5 'Tables'. This is an output file which contains the optional cross tabulations of the dependent variable against each predictor at each node. Depending on the detail option selected, the tables are shown before the grouping, after the grouping, or both. Even if the control deck does not call for the detailed table output, you are required to provide a file name for the tables.

6 'Split rule table'. This file written by CATFIRM is useful for programs that apply the dendrogram to the calibration cases or to future cases. The file contains one line for each node that is split in the analysis. The first number in the line is the node number. The second is the number of the predictor that was used to split that node. The remaining numbers show to which descendant node the cases with each value of that predictor go, starting from the lowest category of the predictor and going up to the highest. The format of the file is perhaps best illustrated by an example. Imagine that node 10 were split on the third predictor, and that the range of values third predictor was 3, 4, 5, 6, 7 and 8, and that the cases where this predictor was 3, 5 or 7 defined descendant node 17, those with the predictor 4 defined descendant node 18, and those with the predictor 6 or 8 defined descendant node 19. (Clearly this predictor is FREE). Then the split rule table would have a line

10 3 17 18 17 19 17 19.

CONFIRM calls for only five files. These are:-

1 'Control deck'. This is the control deck created by DECK, and is input.

2 'Data'. This is the input data file.

3 'Summary'. This is an output file produced by CONFIRM. It contains a section for each node analyzed, showing (i) a list of the predictors with the groupings of the categories of each, (ii) the multiple comparison and Bonferroni significance level of the split on that predictor, (iii) which predictor (if any) is used to split that group, (iv) the one-way analysis of variance of the dependent variable by the grouped levels of the predictor and (v) the numbers, means and standard deviations of the cases in the descendant node.

4 'Split'. This file contains the detail (if requested) of the reduction of each predictor at each node to the final grouping of the categories of that predictor. The output comprises the t statistics used to test the merging of each (simple or composite) category grouping with its neighbors, and the means, before and after grouping, of the cases in each category. Even if the detailed split information was not requested, a file name must be provided in response to this question.

5 'Split rule table'. This file has the same meaning, format and use as the split rule table created by CATFIRM.

Sample outputs

The appendix gives six short extracts of the FIRM outputs to illustrate some of the major features.

CATFIRM Summary file The first sample is from the summary file of the head injuries data, and shows the header information, together with the summary of results on nodes 1, 2 and 3. The header information is self-explanatory, so we will concentrate on the results. In node 1, all the predictors give highly significant splits on the dependent variable. The number of groups selected by CATFIRM varies from 2 to 4:- under the heading 'groups', we see how the categories break down. For example, using EMV would split the cases into 4 groups - EMV=1 and 2 form one group, 3, 4 and 5 a second, ? and 6 a third, and 7 the fourth. The most significant split uses Pupils. It is a binary split, between the pooled class ? or 2,

and class 1. This split has a Bonferroni significance level of 1.8×10^{-19}% and a multiple comparison significance level of 3×10^{-18}%. The smaller of these (both being conservative) is taken as the significance level of the split; this is 1.8×10^{-19}%. The summary table below the split shows how the cases divide between these two - roughly three quarters go to node 2, where the recovery rate is quite variable, and the rest go to node 3, where 90% of the patients are dead or vegetative.

One interesting feature can be seen in the predictors not used for the split - a missing value for MRP and for EMV appear to be predictive, since the missing category in both was found to be most like category 6.

The analysis continues with node number 2. CATFIRM lists the makeup of this node - it is all cases for which Pupils is ? or 2. As the analysis proceeds, this 'makeup' record grows to reflect the successive splits giving rise to the node. In this node, no significant split can be made on Pupils (not surprisingly, since the two classes of Pupils represented in this node were grouped because of their compatibility). All other predictors give significant splits, the significance ranging from 0.2% for Change down to 10^{-11}% for age. Node 2 is split four ways on age, the cut points being at ages 20, 40 and 60. All four nodes are investigated again later in the analysis.

Node 3 represents a homogeneous group with very high mortality. CATFIRM can find no significant splits, and so this node is terminal.

The summary file has a record like these of each node analyzed. Nodes which are too small for analysis however are not reported separately; but their sample sizes and frequency distribution are listed at the point where they are created.

CATFIRM Split file Next, we look at the entries in CATFIRM's split file for the analysis of the first node. There are three slightly different layouts here, illustrated by the monotonic predictor Age, the free predictor EMV, and the floating predictor Change. For the monotonic predictor, only adjacent predictors may be merged. Thus age starts out with 8 groups, giving 7 pairwise χ^2 statistics which are listed after 'Test statistics for grouping'. The smallest of these is 1.3775 for merging categories 2 and 3. This value is well below the merge significance level specified for the run, so these two classes are joined into a

composite, leaving 7 groups. Then merge statistics for these groups are computed, the smallest of which is 1.6163 for merging classes 0 and 1, so these classes are merged. The analysis continues in the same way until 4 composite groups remain. At this stage, the smallest merge statistic - that for merging (23) with (45) - is significant at the 2% level, which is below the run's threshold for merging. Thus no further reduction of the classes of Age occurs.

No more merging being possible, CATFIRM then attempts to split the composite categories. The line 'Test stats for splitting' gives the details of this, showing a X^2 statistic for each possible resplitting point of a composite category. The largest of these statistics is 3.5501, which is far from significant at the 'split' significance level specified for the run, and so no resplitting takes place. Thus (01), (23), (45), (67) is the final grouping of the categories. The 'raw' significance of the resulting 3x4 contingency table is $9.3\times10^{-13}\%$, but the Bonferroni multiplier, which allows for the grouping that went into reducing 8 groups to 4, is 35. The Bonferroni significance level is therefore $35 \times 9.3\times10^{-13}\% = 3.3\times10^{-11}\%$. The multiple comparison significance level is $6.7\times10^{-9}\%$

The floating predictors, as exemplified by EMV, have an extra wrinkle to them. Not only can the monotonic portion of the scale be merged in the same way as with Age, but the floating category can be merged with any of them. Thus the detail starts out with two lines - the test statistics for merging a successive pair on the scale 1-7 on the upper line, and those for merging the floating category with each monotonic class on the second line. CATFIRM checks the full list of (in this case 13) test statistics, and finds that the smallest is 1.0854, for merging classes 4 and 5. This is done and the analysis repeated. In the second stage, it happens that the smallest test statistic is for merging the floating category ? with 6, so these two categories are joined. Once this is done, there is no longer a floating category, just the monotonic categories 1, 2, 3, (45), (?6), and 7, and the subsequent lines of the analysis look like those of a monotonic predictor.

MRP shows a slightly different twist at the last phase. Here the floating category is part of a three-class composite (?67), and so there are three possible binary splits - ? vs (67), (?6) vs 7 and 6 vs (7?). The test statistics for these are shown on two lines of the printout - 2.5 and 4.3 for the first two, and 1.4 for the third. In reading these two lines, one ignores the rightmost ? on the first line and the leftmost on the second.

Again the split statistic is not significant, and so the grouping (12) (345) (?67) is final.

Free predictors with more than three categories involve much more computation than monotonic or floating predictors with the same number of categories. Their analysis (though not the potential computational load) is illustrated by Pupils. Since any two categories may be merged, each step of the merge phase computes and lists the lower triangle of a matrix of pairwise X^2 values - in this case of a 3x3 matrix. The first round of testing finds that the categories ? and 2 are not significantly different, and so merges them. The second shows that the composite class (?2) is very significantly different from class 1, and so the merging ends.

The splitting of grouped categories in a free predictor is done by considering all possible binary splits of every composite class. For each composite, CATFIRM produces a section of analysis listing the classes in the composite (illustrated here by the section 'Test stats for splitting the group: ? 2), and below this, an exhaustive list of the binary splits and the associated test statistics. A logical string of F and T values identifies which categories are grouped together for each such binary split. In this small case, the only possible split is ? against 2, and the single statistic FT 3.6 shows that this split is not significant.

If the splitting phase finds a split that is significant at the 'split' significance level selected for the run, this composite is split, and CATFIRM returns to the first phase of looking for possible merges.

Since a c-category composite class of a free predictor can be resplit in 2^{c-1} ways, testing for resplitting of free predictors with many classes can become an enormous computational burden. It is partly for this reason, and partly because of the method of implementation that CATFIRM as an immutable upper limit of 16 on the number of categories a free predictor may have.

In analyses with free predictors having many categories, a substantial amount of computing time can sometimes be saved with a modest loss in the quality of the final grouping by skipping the resplitting phase. This happens automatically whenever the 'split' significance level specified for the run is larger than the 'merge' significance level.

CATFIRM Table file The other optional file produced on request by CATFIRM is one of the contingency tables before and/or after the grouping of the categories. This is illustrated by the next section of printout, which shows these tables for the first node. Scanning these tables is often helpful in gaining a better perspective to the grouping that went before; in particular of the numbers of cases in the groups that were merged. For example, we made the comment that missingness of MRP and EMV seemed to be informative. This point is supported by the 'before' contingency tables of the outcome against these variables. While the summary table showed that ? was merged with 6, and the split table quantified this with X^2 values, the table output shows that the values 6 and 7 for both these measures are above average, the bulk of the data being around 4, and that the ? category has a considerably better prognosis than average.

CATFIRM Split rule table. The final file is the split rule table. On occasion, some categories of a predictor 'go dead' - although the category was represented in the original full sample, it is empty in the current node. When this occurs, FIRM has no basis for deciding to which descendant node future cases with that value should be assigned. This is signaled by CATFIRM's giving the category the destination node 999, node 999 being the catchall node for such dead values of a predictor used for splitting.

CONFIRM output files

The files produced by CONFIRM closely parallel those of CATFIRM, and so can be dealt with more briefly.

CONFIRM Summary file The summary file shows the splits actually made, and the grouping on predictors which were not used for the splitting. As a matter of interest, while there is no particular reason why they should be, in the first node the predictor groupings and ranking by overall significance of CONFIRM and CATFIRM are quite similar. All predictors are highly significant, but MRP edges out Pupils for the most significant split. The summary file gives the one-way analysis of variance for the split actually made, and then lists the summary statistics of the descendant groups formed by the split. After doing this for node 2, the program notes that descendant group 5 (nearly all of whose members end up dead or vegetative) is too homogeneous to be considered for further partitioning.

CONFIRM Split file In CONFIRM, a single file contains the information that in CATFIRM is split between the Split and the Table files. The listing starts with the summary statistics of the grouping of the cases in that node by the different classes of the predictor. This is followed by the corresponding one-way anova. Next is the printout of the 'merge' section of the analysis. This is considerably more cryptic than the corresponding CATFIRM printout. For a monotonic predictor like Age, the number of merge statistics is initially one fewer than the number of classes, and Student's t statistics are listed in order. For example, the t value for merging classes 0 and 1 is 1.2; that for merging 1 and 2 is -2.3; ... that for merging 6 and 7 is -2.2. The smallest t value is -0.3, for merging classes 2 and 3. This is done, and the number of classes for merging becomes 7, with 6 possible mergings and their associated t values. These t values are listed on the second 'merge stats' line, which shows that the least significantly different mergable pair is 4 and 5, with a t value of 0.6. This merge in turn takes place, as this t value is not significant at the 'merge' significance level selected for the run. This reduction continues until at the final line, both the Student's t values (for merging the composite categories (0123), (45), and (67)) are significant, and the merge testing stops.

A slightly different format is used for a floating predictor, as illustrated by the predictor MRP. Here, while the floating category is on its own, there are two lines of statistics for each stage - the first for merging ? with 1, 1 with 2, 2 with 3 ...; and the second for merging ? with 1, ? with 2, ? with 3 .... At the first merge phase, these lines show that the least significant difference is for categories 3 with 4, and so these categories are merged. The second stage merges 6 with 7 and the third 1 with 2. At the fourth stage, the smallest t is for merging ? with the composite category (67), and after this is done, ? no longer floats, and so for the last two stages these is only a single line of 'merge stats' output.

The final type of printout is that for a free predictor (like Pupils). This is much simpler than in the corresponding CATFIRM case. Here the first stage of the analysis is to sort the categories of the predictor into ascending order of their mean values of the dependent variable. Thereafter, the analysis proceeds just like that of a monotonic predictor in these re-ordered categories.

CONFIRM split rule table The format of this table is identical to that of the corresponding CATFIRM file.

Versions and resources needed

The FIRM codes are made available in three executable forms - two for IBM PC's and compatibles, and the third for Apple Mac Plus and higher.

One IBM variant is for models with 80286 and higher chips and a math coprocessor, while the other will run on basic 8086 machines although its execution times on such machines are not stellar. The 8086 codes have been kept below 200K in size.

Mac enthusiasts should note that the Mac version uses the identical source code to the IBM version, and does not make any use of the Mac interface. It requires that the needed file names be typed in from the keyboard, and not selected with the mouse.

In addition to main storage, the programs require scratch space on disk. In a problem with N cases and M predictors, CATFIRM requires disk space for a scratch file of MN bytes, while CONFIRM requires 3N(M+12) bytes. Ideally, these scratch files should be on RAM disks or hard disks, though floppy disks will work, albeit appreciably more slowly.

Timings were obtained for the analysis of the head injuries data on three different computers - an Apple Mac Plus with hard disk, an IBM PC AT with a coprocessor and hard disk, and a vintage (4.77MHz, 256K) IBM PC with two floppy drives as the sole storage medium. The times taken for the analyses were as follows:-

	CATFIRM	CONFIRM
IBM PC AT	58 seconds	172 seconds
Apple Mac	152 seconds	417 seconds
Basic PC	210 seconds	22 minutes

About half the time taken by the head injuries data was involved in data manipulation (which is an order MN activity) and the remainder on the reduction of categories (which is an order M activity). If we regard one hour as the maximum execution time that can comfortably be tolerated, this means that on an AT problems 20 to 50 times as big as the head injuries data can reasonably be run under CONFIRM and CATFIRM respectively. On the Mac Plus, this range would be 10 to 20 times the size of the head injuries data. Even on the quite minimal

4.77MHz PC with only floppy drives, quite large problems are tractable, particularly with CATFIRM.

While the FIRM codes are running, they give brief progress reports on the screen, showing which node is currently being analyzed and what splits are made. This is done primarily to reassure the user that the programs are doing something and not hung, but particularly at the early stages of analyzing a data set there may be some interest in interrupting the analysis once a handful of nodes have been created to see how the analysis looks.

Apart from that implied by available disk space, none of the versions has any limit on the number of cases. There are however limits on the number of predictors, and on the number of categories the variables may have. At the time of writing, the AT and Mac versions have a limit of 100 predictors, and 20 variable categories in CONFIRM, 16 in CATFIRM. The basic AT version reduces these figures to keep the code small.

Further reading

The chapter by Hawkins and Kass (1982), and the papers discussed there, provide more detail on the ideas and implementations of FIRM. An example of the use of a predecessor to CATFIRM to a much larger data set than the head injuries data set is given by Hooton et al. (1981).

References

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., (1984), 'Classification and Regression Trees', Wadsworth, Belmont.

Hawkins, D. M., and Kass, G. V., (1982), 'Automatic Interaction Detection' in Topics in Applied Multivariate Analysis', ed D M Hawkins, Cambridge University Press.

Heymann, C., (1981), 'XAID - an Extended Automatic Interaction Detector', Internal Report SWISK 28, Council for Scientific and Industrial Research, Pretoria, South Africa.

Hooton, T. M., Haley, R. W., Culver, D. H., White, J. W., Morgan, W. M., and Carroll, R. J., (1981), 'The joint associations of multiple risk factors with the occurrence of nosocomial infections', American Journal of Medicine, 70, 960-970.

Kass, G. V., (1980), 'An exploratory technique for investigating large quantities of categorical data', Applied Statistics, 29, 119-127.

Kass, G. V., (1975), 'Significance testing in, and an extension to

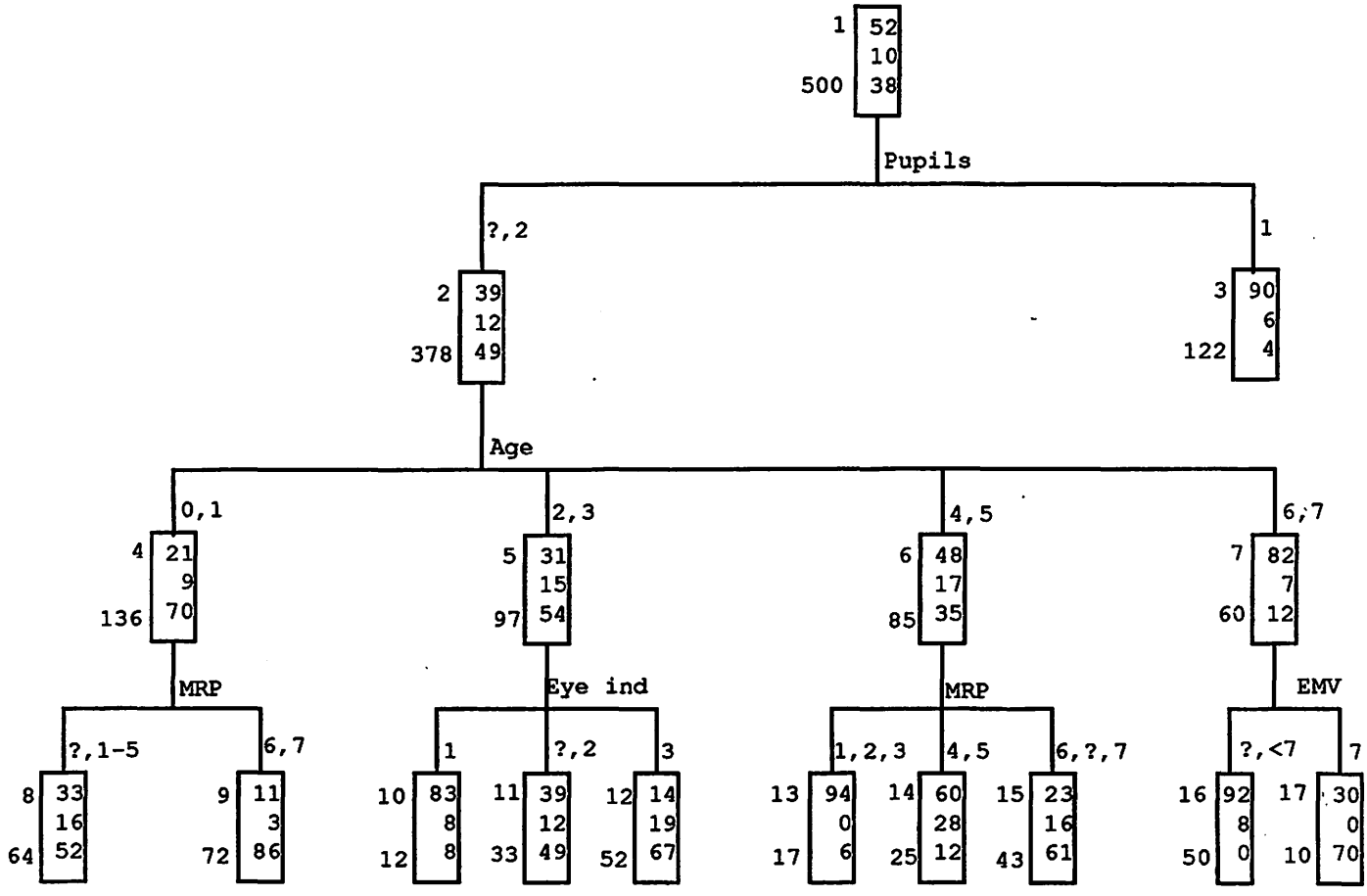
Automatic Interaction Detection', PhD Thesis, University of the Witwatersrand, Johannesburg.

Loh, W-Y., and Vanichsetakul, N., (1988), 'Tree structured classification via generalized discriminant analysis', Journal of the American Statistical Association, 83, 715-725.

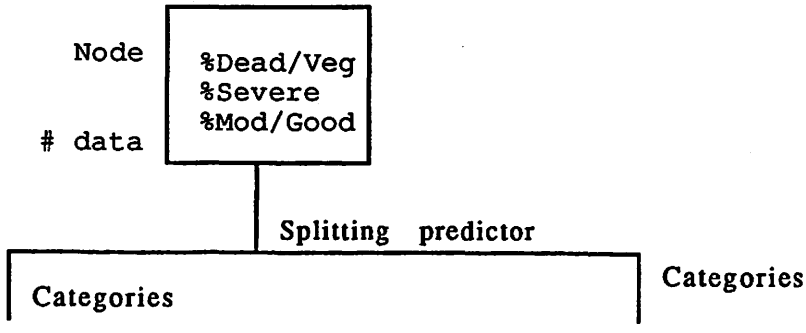
Morgan, J. A., and Sonquist, J. N., (1963), 'Problems in the analysis of survey data: and a proposal', Journal of the American Statistical Association, 58, 415-434.

Titterington, D. M., Murray, G. D., Murray, L. S., Spiegelhalter, D. J., Skene, A. M., Habbema, J. D. F., and Gelpke, G. J., (1981), 'Comparison of discrimination techniques applied to a complex data set of head injured patients', Journal of the Royal Statistical Society, A144, 145-161.

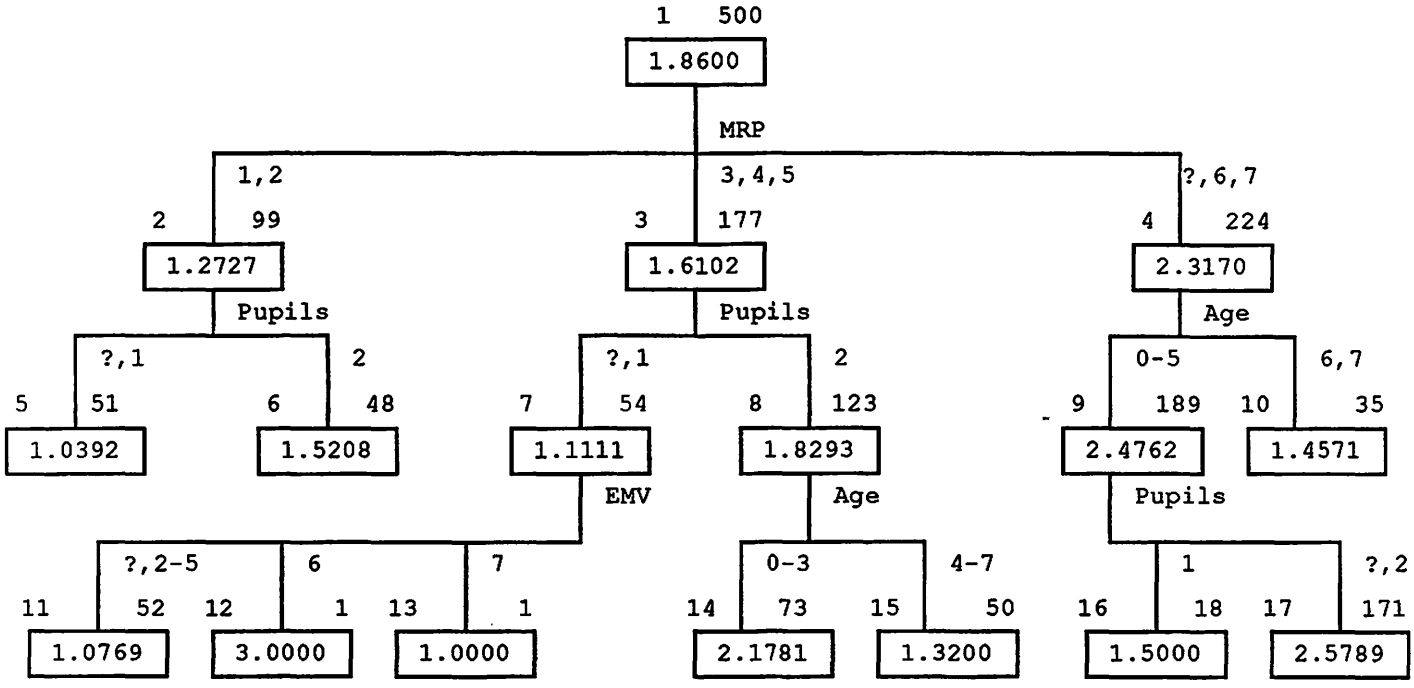
Dendrogram of CATFIRM analysis of head injuries data set



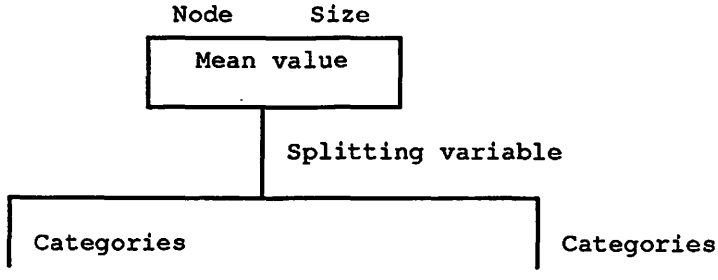
Legend



Dendrogram of CONFIRM analysis of head injuries data



Legend



Printout 1. CATFIRM Summary file

```
                CATFIRM  Formal Inference-based Recursive Modeling.
                Categorical dependent variable

                Copyright  1990  Douglas M Hawkins
                Applied Statistics
                University of Minnesota
Program dimensions
Maximum number of predictors            100
Maximum number of categories in
        Predictors                       16
        Dependent variable               16

Outcome                             has   3 categories called:
Dead/veg
Severe
Good
There are  6 predictors as follows
Type   No. cats   Cat symbols   Use?  Split% Merge%
Mono        8  01234567          May     4.90   5.00    age
Float 1     8  ?1234567          May     4.90   5.00    EMV
Float 1     8  ?1234567          May     4.90   5.00    MRP
Float 1     4  ?123              May     4.90   5.00    Change
Float 1     4  ?123              May     4.90   5.00    Eye ind
Free        3  ?12               May     4.90   5.00    Pupils
Option  1 is       0.
Option  2 is    1000.
Option  3 is      50.
Option  4 is       1.
Option  5 is       1.
Option  6 is      30.
Option  7 is       0.
Option  8 is       0.
Options in effect:
Tables printed  as column percentages
Detail output on file split
Tables given before&after each step
To be analysed, a group must:
        have at least   50 cases;
        be significant at the   .500% level;
        be Bonferroni significant at the   .500% level.
The run will terminate when    30 groups have been formed.
Standard Pearson X^2 statistic is used
```

```
Summary of results of node number     1 predecessor node number    0
Total group
no.  Name                Signif %   Bonf sig %    MC sig %   groups
  1 age               9.3157E-13%  3.2605E-11%  6.6815E-09%  4 01 23 45 67
  2 EMV               3.9659E-20%  3.7677E-18%  1.2262E-15%  4 12 345 ?6 7
  3 MRP               3.4067E-20%  1.7374E-18%  2.2721E-14%  3 12 345 ?67
  4 Change              .0125102%    .0625510%    .6300625%  2 ?1 23
  5 Eye ind           6.1672E-19%  3.0836E-18%  1.5879E-17%  3 1 ?2 3
  6 Pupils            6.0487E-20%  1.8146E-19%  3.0157E-18%  2 ?2 1

Characteristics of the best predictor
  6 Pupils            6.0487E-20%  1.8146E-19%  3.0157E-18%  2 ?2 1
**************************************************************************
predictor   6   Pupils                *percent*  total number    500
                   ?,2     1   Total
Dead/veg            39.4  90.2  51.8
Severe              11.9   5.7  10.4
Good                48.7   4.1  37.8
  totals (100%)      378   122   500
        Raw significance of table is  6.0487E-20%

 The descendant nodes are numbered   2   3

Summary of results of node number     2 predecessor node number    1
Makeup   Pupils    (?,2)
no.  Name                Signif %   Bonf sig %    MC sig %   groups
  1 age               3.6963E-13%  1.2937E-11%  2.9172E-09%  4 01 23 45 67
  2 EMV               9.4560E-10%  4.8226E-08%  3.1324E-05%  3 12 ?345 67
  3 MRP               1.1675E-09%  5.9544E-08%  3.7290E-05%  3 12345 ?6 7
  4 Change              .0393407%    .1967034%   1.5570522%  2 ?1 23
  5 Eye ind           2.4476E-07%  1.2238E-06%  2.9430E-06%  3 1 ?2 3
  6 Pupils              100.0000%    100.0000%    100.0000%  1 ?2

Characteristics of the best predictor
  1 age               3.6963E-13%  1.2937E-11%  2.9172E-09%  4 01 23 45 67
**************************************************************************
predictor   1   age                   *percent*  total number    378
                  0,1     2,3     4,5     6,7   Total
Dead/veg           21.3    30.9    48.2    81.7  39.4
Severe              8.8    15.5    16.5     6.7  11.9
Good               69.9    53.6    35.3    11.7  48.7
  totals (100%)     136      97      85      60   378
        Raw significance of table is  3.6963E-13%

 The descendant nodes are numbered   4   5   6   7

Summary of results of node number     3 predecessor node number    1
Makeup   Pupils    (1)
no.  Name                Signif %   Bonf sig %    MC sig %   groups
  1 age               100.0000%    100.0000%    100.0000%  1 01234567
  2 EMV               .0470654%   2.4003330%  12.6084000%  3 12 345 ?67
  3 MRP               .0994236%   1.2925063%  46.2672340%  2 123456 ?7
  4 Change            100.0000%    100.0000%    100.0000%  1 ?123
  5 Eye ind           .3083922%   1.5419610%   7.2455911%  2 12 ?3
  6 Pupils            100.0000%    100.0000%    100.0000%  1 1

Characteristics of the best predictor
  3 MRP               .0994236%   1.2925063%  46.2672340%  2 123456 ?7
This predictor is not significant
```

Printout 2. CATFIRM Split file

```
   1Total group
************************************************************************
Monotonic age
Table has chi-square   86.310, with df  14 and significance 1.8797E-10%
  8 groups:                    0     1     2     3     4     5     6     7
Test statistics for grouping:     1.6   9.8   1.4   4.0   1.7   7.7   5.1

Min stat is  1.3775, to merge (2) and (3).       d.f.   2, sig 50.2210730%
  7 groups:                    0     1    (2     3)    4     5     6     7
Test statistics for grouping:     1.6  10.7         7.9   1.7   7.7   5.1

Min stat is  1.6163, to merge (0) and (1).       d.f.   2, sig 44.5691000%
  6 groups:                   (0     1)   (2     3)    4     5     6     7
Test statistics for grouping:           9.5         7.9   1.7   7.7   5.1

Min stat is  1.7396, to merge (4) and (5).       d.f.   2, sig 41.9042620%
  5 groups:                   (0     1)   (2     3)   (4     5)    6     7
Test statistics for grouping:           9.5         7.8         7.1   5.1

Min stat is  5.0593, to merge (6) and (7).       d.f.   2, sig  7.9685340%
  4 groups:                   (0     1)   (2     3)   (4     5)   (6     7)
Test statistics for grouping:           9.5         7.8        14.4

Min stat is  7.7518, to merge (23) and (45).     d.f.   2, sig  2.0735931%
  4 groups:             (0     1) (2     3) (4     5) (6     7)
Test stats for splitting
                          1.6       1.4       1.7       5.1
Max stat is  5.0593 to split group (67). d.f.   2 significance  7.9685340%
Best is  4 groups, with chi square   77.984 d.f.   6
Bonferroni multiplier, raw significance and MC significance
     35.                      9.3157E-13%         6.6815E-09%
************************************************************************
************************************************************************
Float EMV
Table has chi-square  120.818, with df  14 and significance 4.3503E-17%
  8 groups:                    1     2     3     4     5     6     7
Test statistics for grouping:     2.6   9.4   3.4   1.1   9.4   6.5
and for grouping ? with       18.1  24.7   5.8   2.0   1.4   1.6   9.4
Min stat is  1.0854, to merge (4) and (5).       d.f.   2, sig 58.1174310%
  7 groups:                    1     2     3    (4     5)    6     7
Test statistics for grouping:     2.6   9.4   3.6        11.5   6.5
and for grouping ? with       18.1  24.7   5.8   1.8   1.6   9.4
Min stat is  1.5590, to merge (?) and (6).       d.f.   2, sig 45.8631240%
  6 groups:                    1     2     3    (4     5)   (?     6)    7
Test statistics for grouping:     2.6   9.4   3.6        10.7         9.6

Min stat is  2.6283, to merge (1) and (2).       d.f.   2, sig 26.8698100%
  5 groups:                   (1     2)    3    (4     5)   (?     6)    7
Test statistics for grouping:          14.6   3.6        10.7         9.6

Min stat is  3.5501, to merge (3) and (45).      d.f.   2, sig 16.9478200%
  4 groups:                   (1     2)   (3     4     5)   (?     6)    7
Test statistics for grouping:          35.3              15.1         9.6

Min stat is  9.5637, to merge (?6) and (7).      d.f.   2, sig   .8380363%
  4 groups:             (1     2) (3     4     5) (?     6)   7
Test stats for splitting
                        2.6       3.6   1.3       1.6
Max stat is  3.5501 to split group (345). d.f.   2 significance 16.9478340%
```

```
Best is  4 groups, with chi square  113.393 d.f.   6
Bonferroni multiplier, raw significance and MC significance
     95.                      3.9659E-20%          1.2262E-15%
*****************************************************************************
*****************************************************************************
Float MRP
Table has chi-square  117.765, with df  14 and significance 1.7216E-16%
  8 groups:                    1     2     3     4     5     6     7
Test statistics for grouping:      .7   3.8   1.9   3.7   5.0   3.0
and for grouping ? with       17.9  16.6   6.1   8.8   1.4    .9   4.5
Min stat is   .7021, to merge (1) and (2).       d.f.   2, sig 70.3938900%
  7 groups:                   (1     2)    3     4     5     6     7
Test statistics for grouping:          5.9   1.9   3.7   5.0   3.0
and for grouping ? with       23.1   6.1   8.8   1.4    .9   4.5
Min stat is   .8839, to merge (?) and (6).       d.f.   2, sig 64.2795710%
  6 groups:                   (1     2)    3     4     5    (?     6)    7
Test statistics for grouping:          5.9   1.9   3.7   4.7         4.3

Min stat is  1.9242, to merge (3) and (4).       d.f.   2, sig 38.2081100%
  5 groups:                   (1     2)   (3     4)    5    (?     6)    7
Test statistics for grouping:          8.6         3.3   4.7         4.3

Min stat is  3.3063, to merge (34) and (5).      d.f.   2, sig 19.1445600%
  4 groups:                   (1     2)   (3     4     5)   (?     6)    7
Test statistics for grouping:         11.8               34.9         4.3

Min stat is  4.2724, to merge (?6) and (7).      d.f.   2, sig 11.8101740%
  3 groups:                   (1     2)   (3     4     5)   (?     6     7)
Test statistics for grouping:         11.8               56.1

Min stat is 11.8477, to merge (12) and (345).    d.f.   2, sig   .2674831%
  3 groups:           (1     2) (3     4     5) (?     6     7    ?)
Test stats for splitting
                          .7       1.5   3.3       2.5   4.3
                                                              1.4
Max stat is  4.2724 to split group (?67). d.f.   2 significance 11.8101740%
Best is  3 groups, with chi square  106.856 d.f.   4
Bonferroni multiplier, raw significance and MC significance
     51.                      3.4067E-20%          2.2721E-14%
*****************************************************************************
*****************************************************************************
Float Change
Table has chi-square   25.981, with df   6 and significance   .0224459%
  4 groups:                    1     2     3
Test statistics for grouping:     12.9   3.7
and for grouping ? with         3.9   4.0   6.3
Min stat is  3.6648, to merge (2) and (3).       d.f.   2, sig 16.0026400%
  3 groups:                    1    (2     3)
Test statistics for grouping:     21.0
and for grouping ? with         3.9   5.9
Min stat is  3.9054, to merge (?) and (1).       d.f.   2, sig 14.1887530%
  2 groups:                   (?     1)   (2     3)
Test statistics for grouping:          18.0

Min stat is 17.9728, to merge (?1) and (23).     d.f.   2, sig   .0125102%
  2 groups:           (?     1) (2     3)
Test stats for splitting
                           3.9       3.7
Max stat is  3.9054 to split group (?1). d.f.   2 significance 14.1887700%
Best is  2 groups, with chi square   17.973 d.f.   2
Bonferroni multiplier, raw significance and MC significance
      5.                       .0125102%           .6300625%
```

```
**************************************************************************
**************************************************************************
Float Eye ind
Table has chi-square  105.257, with df   6 and significance 2.0024E-18%
  4 groups:                    1     2     3
Test statistics for grouping:    30.3  17.5
and for grouping ? with       48.4   4.6  12.3
Min stat is  4.6056, to merge (?) and (2).       d.f.   2, sig  9.9976091%
  3 groups:                    1    (?     2)   3
Test statistics for grouping:    45.9        19.9

Min stat is 19.9100, to merge (?2) and (3).      d.f.   2, sig   .0047490%
  3 groups:              1   (?     2)   3
Test stats for splitting
                              4.6
Max stat is  4.6056 to split group (?2). d.f.   2 significance  9.9976160%
Best is  3 groups, with chi square  100.952 d.f.   4
Bonferroni multiplier, raw significance and MC significance
      5.                     6.1672E-19%          1.5879E-17%
**************************************************************************
**************************************************************************
Free Pupils
Table has chi-square  101.477 with d.f.   4 and signif 4.7688E-19%
    Merge phase    3 groups        ?     1
                          1         9.9
                          2         3.6  99.9
Min stat  3.62 to merge groups  1,  3 d.f.   2  signif 16.3588840%
    Merge phase    2 groups        ?,2
                          1         97.7
Min stat 97.71 to merge groups  1,  2 d.f.   2  signif 6.0487E-20%
Split phase    2           groups: (?    2 ) (1)
Test stats for splitting the group: ? 2
FT   3.6
FT   3.6
Max stat is  3.6208 to split group numbered  1 as                1
 d.f.   2 significance 16.3588920%
Best is  2 groups with chi squared   97.714 d.f.   2
 Bonferroni multiplier, raw significance and MC significance
      3.                     6.0487E-20%          3.0157E-18%
**************************************************************************
Best to use var  6 Pupils              to give 2 new groups
```

Printout 3. CATFIRM Table file

```
   1Total group  before grouping
*************************************************************************
predictor   1   age                  *percent*  total number    500
                    0     1     2     3     4     5     6     7   Total
Dead/veg          40.0  31.5  41.6  47.3  65.6  56.9  80.3 100.0  51.8
Severe             9.1   7.2  18.2  10.9   8.2  15.5   8.2    .0  10.4
Good              50.9  61.3  40.3  41.8  26.2  27.6  11.5    .0  37.8
  totals (100%)     55   111    77    55    61    58    61    22   500
        Raw significance of table is  1.8797E-10%

    1Total group  before grouping
*************************************************************************
predictor   2   EMV                  *percent*  total number    500
                    ?     1     2     3     4     5     6     7   Total
Dead/veg          39.3 100.0  87.5  63.5  54.1  50.0  33.8  15.4  51.8
Severe            14.3    .0   6.3  15.4  10.8  15.6   7.7   6.2  10.4
Good              46.4    .0   6.3  21.2  35.1  34.4  58.5  78.5  37.8
  totals (100%)     28    19    64    52   111    96    65    65   500
        Raw significance of table is  4.3503E-17%

    1Total group  before grouping
*************************************************************************
predictor   3   MRP                  *percent*  total number    500
                    ?     1     2     3     4     5     6     7   Total
Dead/veg          33.3  86.8  80.3  66.7  65.8  50.0  28.6  28.6  51.8
Severe            19.0   5.3   8.2   6.1  14.0  13.3  13.2   6.3  10.4
Good              47.6   7.9  11.5  27.3  20.2  36.7  58.2  65.2  37.8
  totals (100%)     21    38    61    33   114    30    91   112   500
        Raw significance of table is  1.7216E-16%

    1Total group  before grouping
*************************************************************************
predictor   4   Change               *percent*  total number    500
                    ?     1     2     3   Total
Dead/veg          54.5  65.7  43.5  39.1  51.8
Severe             9.1   8.4  15.7   9.1  10.4
Good              36.4  25.9  40.9  51.8  37.8
  totals (100%)    132   143   115   110   500
        Raw significance of table is     .0224459%

    1Total group  before grouping
*************************************************************************
predictor   5   Eye ind              *percent*  total number    500
                    ?     1     2     3   Total
Dead/veg          50.9  92.7  58.9  32.1  51.8
Severe             6.4   5.2  12.3  14.0  10.4
Good              42.7   2.1  28.8  53.8  37.8
  totals (100%)    110    96    73   221   500
        Raw significance of table is  2.0024E-18%
   1Total group  before grouping
*************************************************************************
predictor   6   Pupils               *percent*  total number    500
                    ?     1     2   Total
Dead/veg          61.5  90.2  38.6  51.8
Severe            15.4   5.7  11.8  10.4
Good              23.1   4.1  49.6  37.8
  totals (100%)     13   122   365   500
        Raw significance of table is  4.7688E-19%
```

Total group
1Total group after grouping

predictor 1 age *percent* total number 500

	0,1	2,3	4,5	6,7	Total
Dead/veg	34.3	43.9	61.3	85.5	51.8
Severe	7.8	15.2	11.8	6.0	10.4
Good	57.8	40.9	26.9	8.4	37.8
totals (100%)	166	132	119	83	500

Raw significance of table is 9.3157E-13%

1Total group after grouping

predictor 2 EMV *percent* total number 500

	1,2	3,4,5	?,6	7	Total
Dead/veg	90.4	54.4	35.5	15.4	51.8
Severe	4.8	13.5	9.7	6.2	10.4
Good	4.8	32.0	54.8	78.5	37.8
totals (100%)	83	259	93	65	500

Raw significance of table is 3.9659E-20%

1Total group after grouping

predictor 3 MRP *percent* total number 500

	1,2	3,4,5	?,6,7	Total
Dead/veg	82.8	63.3	29.0	51.8
Severe	7.1	12.4	10.3	10.4
Good	10.1	24.3	60.7	37.8
totals (100%)	99	177	224	500

Raw significance of table is 3.4067E-20%

1Total group after grouping

predictor 4 Change *percent* total number 500

	?,1	2,3	Total
Dead/veg	60.4	41.3	51.8
Severe	8.7	12.4	10.4
Good	30.9	46.2	37.8
totals (100%)	275	225	500

Raw significance of table is .0125102%

1Total group after grouping

predictor 5 Eye ind *percent* total number 500

	1	?,2	3	Total
Dead/veg	92.7	54.1	32.1	51.8
Severe	5.2	8.7	14.0	10.4
Good	2.1	37.2	53.8	37.8
totals (100%)	96	183	221	500

Raw significance of table is 6.1672E-19%

1Total group after grouping

predictor 6 Pupils *percent* total number 500

	?,2	1	Total
Dead/veg	39.4	90.2	51.8
Severe	11.9	5.7	10.4
Good	48.7	4.1	37.8
totals (100%)	378	122	500

Raw significance of table is 6.0487E-20%

Printout 4. CATFIRM Split Rule file

```
1   6   2   3   2
2   1   4   4   5   5   6   6   7   7
4   3   8   8   8   8   8   8   9   9
5   5  11  10  11  12
6   3  15  13  13  13  14  14  15  15
7   2  16 999  16  16  16  16  16  17
```

Printout 5. CONFIRM summary file

```
                CONFIRM  Formal Inference-based Recursive Modeling
                Continuous dependent variable

                Copyright 1990  Douglas M Hawkins
                Applied Statistics
                University of Minnesota

    Dependent variable no.     1 name=Outcome
    Predictor variables
no  posn name              no of cats    split  merge  may?     type
 1  2    age                        8    5.000  4.900  yes      mono
 2  3    EMV                        8    5.000  4.900  yes      flt
 3  4    MRP                        8    5.000  4.900  yes      flt
 4  5    Change                     4    5.000  4.900  yes      flt
 5  6    Eye ind                    4    5.000  4.900  yes      flt
 6  7    Pupils                     3    5.000  4.900  yes      free
Run options in effect
Full split/merge details of predictors
For a group to be analyzed, it must:-
   contain at least    40 cases;
   have at least proportion  .01000 of starting ssd.
Minimum % raw significance for a split     .100
Minimum % Bonferroni significance for a split     .500
Analysis will stop after   30 groups have been formed
Error variance is two-group

*************************************************************************
    Analysis of group no.  1          previous group no.  0
 no.  name                        mc-sig(%)    bon-f-sig(%) grouping
  1   age                        1.5296E-10   1.1490E-12   0123/45/67
  2   EMV                        1.7011E-20   1.313E-021   12/345/?6/7
  3   MRP                        4.1531E-20   1.929E-022   12/345/?67
  4   Change                       .0769788     .0201893   ?1/23
  5   Eye ind                    5.711E-021   3.250E-021   1/?2/3
  6   Pupils                     1.232E-021   2.694E-022   1/?2

    Best predictor
  3  MRP                         4.1531E-20   1.929E-022   12/345/?67

    Analysis of variance
                Sum of squares      mean square      degrees of freedom
    Grouping           91.9664          45.9832                    2
    Error             346.2336            .6966                  497

    F-value      =         66.006
    significance=  0.378E-23
    bonf -sig.   =  0.193E-21
    mc-sig.      =  0.415E-19
    conserv.sig.=  0.193E-21
    Grouping is significant at the   0.193E-21:- level(conservative)

Statistics for grouping
Node          Mean       s.d.     size      s.e. (mean)
   2        1.2727      .6360       99      .06392
   3        1.6102      .8531      177      .06413
   4        2.3170      .8947      224      .05978
```

```
    Analysis of group no.  2          previous group no.  1
 no.  name                          mc-sig(%)    bon-f-sig(%) grouping
  1  age                            5.6363001     .1168782    0123/4567
  2  EMV                            1.7234762     .0759155    ?12/3456
  3  MRP                             100.0000     100.0000    12
  4  Change                          .1766159     .0478362    ?/123
  5  Eye ind                         .2085087     .0577924    1/?23
  6  Pupils                          .0551486     .0308304    ?1/2

    Best predictor
  6  Pupils                          .0551486     .0308304    ?1/2

    Analysis of variance
                Sum of squares      mean square      degrees of freedom
    Grouping          5.7356          5.7356                  1
    Error            33.9007           .3495                 97

    F-value      =         16.411
    significance=  0.103E-01
    bonf -sig.   =  0.308E-01
    mc-sig.      =  0.551E-01
    conserv.sig.=  0.308E-01
    Grouping is significant at the   0.308E-01:- level(conservative)

Statistics for grouping
Node          Mean      s.d.     size      s.e. (mean)
   5        1.0392      .2801      51      .03922
   6        1.5208      .7987      48      .11528

   1 of these groups fail(s) thresholds for further analysis.
Group    5 not enough SSD for further  analysis.  SSD is       3.921569

*************************************************************************
    Analysis of group no.  3          previous group no.  1
 no.  name                          mc-sig(%)    bon-f-sig(%) grouping
  1  age                             .2892236     .0018524    0123/4567
  2  EMV                             100.0000     100.0000    ?234567
  3  MRP                             100.0000     100.0000    345
  4  Change                          100.0000     100.0000    ?123
  5  Eye ind                         .0181600     .0039246    1/?23
  6  Pupils                        6.4769E-05   2.6932E-05    ?1/2

    Best predictor
  6  Pupils                        6.4769E-05   2.6932E-05    ?1/2

    Analysis of variance
                Sum of squares      mean square      degrees of freedom
    Grouping         19.3537         19.3537                  1
    Error           108.7480           .6214                175

    F-value      =         31.145
    significance=  0.898E-05
    bonf -sig.   =  0.269E-04
    mc-sig.      =  0.648E-04
    conserv.sig.=  0.269E-04
    Grouping is significant at the   0.269E-04:- level(conservative)

Statistics for grouping
Node          Mean      s.d.     size      s.e. (mean)
   7        1.1111      .3720      54      .05062
   8        1.8293      .9117     123      .08221
```

Printout 6 CONFIRM Split file

```
*******************************************************************************
    Analysis of group no.  1     previous group no.  0     mean=     1.860
size=  500
    predictor no.  1   age
    statistics before merging
 Cate      0       1       2       3       4       5       6       7
 mean     2.11    2.30    1.99    1.95    1.61    1.71    1.31    1.00
 size     55     111      77      55      61      58      61      22
 Anova  sse      dfe      ssh     dfh   r-square    f        sign(%)
      372.015    492    66.185     7      .1510    12.5046  0.8757E-12
 merge stats  1.2 -2.3  -.3 -2.0   .6 -2.8 -2.2
 merge stats  1.2 -2.8 -2.6   .6 -2.8 -2.2
 merge stats  1.2 -2.8 -2.8 -2.7 -2.2
 merge stats -2.4 -2.8 -2.7 -2.2
 merge stats -2.4 -2.8 -3.9
 merge stats -4.6 -3.9

    statistics after merging
 Cate      0123   45      67
 mean     2.12    1.66    1.23
 size    298     119      83
 Anova  sse      dfe      ssh     dfh   r-square    f        sign(%)
      380.414    497    57.786     2      .1319    37.7480  0.5471E-13
    predictor no.  2   EMV
    statistics before merging
 Cate      ?       1       2       3       4       5       6       7
 mean     2.07    1.00    1.19    1.58    1.81    1.84    2.25    2.63
 size     28      19      64      52     111      96      65      65
 Anova  sse      dfe      ssh     dfh   r-square    f        sign(%)
      341.183    492    97.017     7      .2214    19.9861  0.1274E-20
 merge stats -5.0  1.5  3.1  1.6   .3  2.7  2.6
             -5.0 -5.7 -2.4 -1.3 -1.2   .8  3.1
 merge stats -5.0  1.5  3.1  1.8  3.2  2.6
             -5.0 -5.7 -2.4 -1.3   .8  3.1
 merge stats  1.5  3.1  1.8  3.2  3.1
 merge stats  3.9  1.8  3.2  3.1
 merge stats  6.1  3.8  3.1

    statistics after merging
 Cate      12      345     ?6      7
 mean     1.14    1.78    2.19    2.63
 size     83     259      93      65
 Anova  sse      dfe      ssh     dfh   r-square    f        sign(%)
      344.931    496    93.269     3      .2128    44.7058  0.1382E-22
    predictor no.  3   MRP
    statistics before merging
 Cate      ?       1       2       3       4       5       6       7
 mean     2.14    1.21    1.31    1.61    1.54    1.87    2.30    2.37
 size     21      38      61      33     114      30      91     112
 Anova  sse      dfe      ssh     dfh   r-square    f        sign(%)
      342.575    492    95.625     7      .2182    19.6192  0.3351E-20
 merge stats -4.8   .8  1.8  -.4  1.9  2.3   .5
             -4.8 -4.4 -2.1 -3.1 -1.0   .7  1.0
 merge stats -4.8   .8  2.1  1.8  2.3   .5
             -4.8 -4.4 -3.0 -1.0   .7  1.0
 merge stats -4.8   .8  2.1  1.8  2.7
             -4.8 -4.4 -3.0 -1.0   .9
 merge stats -5.2  2.9  1.8  2.7
             -5.2 -3.0 -1.0   .9
```

```
merge stats  2.9  1.8  2.6
merge stats  3.4  8.0

    statistics after merging
Cate      12      345     ?67
mean     1.27    1.61    2.32
size     99     177     224
Anova  sse      dfe      ssh      dfh    r-square     f         sign(%)
     346.234    497     91.966     2      .2099     66.0065   0.3781E-23
    predictor no.  4   Change
    statistics before merging
Cate      ?       1       2       3
mean     1.82    1.60    1.97    2.13
size    132     143     115     110
Anova  sse      dfe      ssh      dfh    r-square     f         sign(%)
     419.056    496     19.144     3      .0437     7.5530   0.5987E-02
merge stats -2.0  3.3  1.2
            -2.0  1.3  2.5
merge stats -2.0  4.6
            -2.0  2.2
merge stats  4.1

    statistics after merging
Cate      ?1      23
mean     1.71    2.05
size    275     225
Anova  sse      dfe      ssh      dfh    r-square     f         sign(%)
     423.604    498     14.596     1      .0333     17.1594   0.4038E-02
    predictor no.  5   Eye ind
    statistics before merging
Cate      ?       1       2       3
mean     1.92    1.09    1.70    2.22
size    110      96      73     221
Anova  sse      dfe      ssh      dfh    r-square     f         sign(%)
     351.364    496     86.836     3      .1982     40.8602   0.1305E-20
merge stats -7.9  6.0  4.3
            -7.9 -1.5  2.8
merge stats  7.4  4.2

    statistics after merging
Cate      1       ?2      3
mean     1.09    1.83    2.22
size     96     183     221
Anova  sse      dfe      ssh      dfh    r-square     f         sign(%)
     353.480    497     84.720     2      .1933     59.5594   0.6500E-21
    predictor no.  6   Pupils
    statistics before merging
Cate      ?       1       2
mean     1.14    1.62    2.11
size    122      13     365
Anova  sse      dfe      ssh      dfh    r-square     f         sign(%)
     351.325    497     86.875     2      .1983     61.4491   0.1422E-21
merge stats  3.2  1.9
merge stats 10.9

    statistics after merging
Cate      1       ?2
mean     1.14    2.09
size    122     378
Anova  sse      dfe      ssh      dfh    r-square     f         sign(%)
     354.390    498     83.810     1      .1913    117.7717   0.8980E-22
```

Printout 7 CONFIRM split rule file

1	3	4	2	2	3	3	3	4	4
2	6	5	5	6					
3	6	7	7	8					
4	1	9	9	9	9	9	9	10	10
7	2	11	999	11	11	11	11	12	13
8	1	14	14	14	14	15	15	15	15
9	6	17	16	17					